REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	31 March 1998	**Technical Report**

4. TITLE AND SUBTITLE
Experimental Evaluation of Loss Perception in Continuous Media

5. FUNDING NUMBERS
DAAH04-96-1-0341

6. AUTHOR(S)
D. Wijesekera, J. Srivastava, A. Nerode, M. Foresti

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Regents of the University of California
c/o Sponsored Projects Office
336 Sproul Hall
Berkeley, CA 94720-5940

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 35873.67-MA-MUR

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by the documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

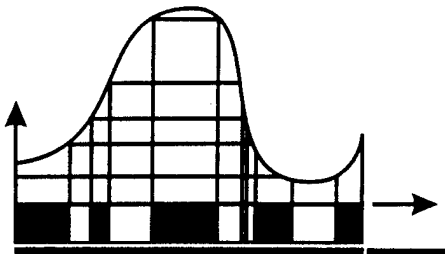
13. ABSTRACT (Maximum 200 words)
Perception of multimedia quality, specified by quality of service metrics can be used by system designers to optimize customer satisfaction within resource bounds enforced by general purpose computing platforms. Media losses, rate variations and transient synchronization losses have been speculated to affect human perception of multimedia quality. This paper presents metrics to measure such defects and results of a series of user experiments that justify such speculations. Results of the study provide bounds on losses, rate variations and transient synchronization losses as a function of user satisfaction, in the form of Likert values. It is shown how these results can be used by algorithm designers of underlying multimedia systems.

14. SUBJECT TERMS
quality of service, user studies, media losses, metrics

15. NUMBER OF PAGES
19

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	**UNCLASSIFIED**	**UNCLASSIFIED**	**UL**



Center for Foundations of Intelligent Systems

Technical Report
98-02

Experimental Evaluation of Loss Perception in Continuous Media

D. Wijesekera, J. Srivastava, A. Nerode and M. Foresti

January 1998

CORNELL
UNIVERSITY

625 Rhodes Hall, Ithaca, NY 14853 (607) 255-8005

Technical Report
98-02

Experimental Evaluation of Loss Perception in Continuous Media

D. Wijesekera, J. Srivastava, A. Nerode
and M. Foresti

January 1998

Experimental Evaluation of Loss Perception in Continuous Media*

Duminda Wijesekera, Jaideep Srivastava, Anil Nerode† and Mark Forestiǂ
Department of Computer Science, University of Minnesota,
Minneapolis, MN 55455.
Institute for Intelligent Machines, Cornell University, Ithaca, NY 14853‡.
Rome Laboratory, Griffis Air Force Base, Rome, NY 14853ǂ
e-mail: {wijesek|srivasta}@cs.umn.edu, forestim@rl.af.mil

Abstract

Perception of multimedia quality, specified by quality of service (QoS) metrics can be used by system designers to optimize customer satisfaction within resource bounds enforced by general purpose computing platforms. Media losses, rate variations and transient synchronization losses have been speculated to affect human perception of multimedia quality. This paper presents metrics to measure such defects, and results of a series of user experiments that justify such speculations. Results of the study provide bounds on losses, rate variations and transient synchronization losses as a function of user satisfaction, in the form of Likert values. It is shown how these results can be used by algorithm designers of underlying multimedia systems.

Keywords: Quality of Service, User Studies, Media Losses, Metrics

1 Introduction

Multimedia systems, characterized by integrated computer-controlled generation, manipulation, presentation, storage and communication of independent discrete and continuous media (CM) data [SGN96], have to compete for the same clientele that has already been accustomed to high standards set by radio and broadcast television. Given the non-deterministic nature of general purpose computing platforms, it is a challenge to provide the high quality of presentations comparable to services such as broadcast TV, which is based on an architecture supported by dedicated cable networks serviced by special purpose tape drives. Fortunately, due to inherent limitations of human perception, some loss of quality can be tolerated. Consequently, it is sufficient to provide multimedia services to be within such tolerable limits, because an application catering for human consumption needs to be good with respect to its human perceptual limitations. Determining such tolerances to errors, commonly referred to as *user studies*, is still in its infancy according to [SGN96, Geo96]. The current paper reports results of some experiments in this area in determining human tolerances to lossy media.

Two widely quoted studies in the area of user studies of multimedia systems include [Ste96, AFKN94]. Based on an extensive study, the former concluded that audio-video lip-synchronization errors of 80 ms. were un-noticed and those up to 120 ms. were detectable but tolerated, and above 120 ms, in-tolerable. For audio-pointer synchronization, these limits were respectively 200 and 1000

*This work is supported by Air Force contract number F30602-96-C-0130 to Honeywell Inc, via subcontract number B09030541/AF to the University of Minnesota, and DOD MURI grant DAAH04-96-10341 to Cornell University

ms. In the latter study, perceptual effects of different frame rates were investigated with respect to audio-visual clips with high temporal, audio and video content.

To the best of our knowledge, both these experiments have been carried out in the presence of loss-less CM streams. During the prototyping and demonstrating phases of a multimedia testbed [HRKHS96], we noticed that missing a few media units does not result in considerable user discontent, provided that not too many media units are missed consecutively, and such misses occur infrequently. We also noticed that our CM streams would drift in and out of synchronization without noticeable user dissatisfaction. Based on these observations, we were inspired to investigate the perceptual tolerance to discontinuity caused by media losses and repetitions, and to that of varying degrees of mis-synchronization. As in the case of pioneering user experiments reported in [Ste96], we designed a mathematical model and metrics of continuity and synchronization, in the presence of media losses [WS96]. This paper reports the results of a user study to validate those metrics and consequently, quantify human tolerance of transient continuity and synchronization losses with respect to audio and video.

Our results indicate that patterns of user sensitivity varies depending on the type of defect. Viewer discontent for aggregate video losses gradually increases with the amount of loss, whereas for other types of losses, mis-synchronizations, there is a sharp rise in user discontent upto a certain value of the defect and then the discontent plateaus out. Rate fluctuations rest somewhere in between, and our experiments indicate that humans are very sensitive to audio losses as compared to video losses. We concluded that 17/100 to 23/100 average video losses are tolerated, and above 23/100 is unacceptable. For audio, although our experiments were inconclusive due to reasons that are discussed later, we concluded that an average of 21/100 silence elimination does not result in user discontent. Furthermore, as observed, a consecutive video loss of about two video frames in 100 does not cause user dissatisfaction. Although losing two consecutive video frames is noticed by most users, once this threshold is reached there is not much room for quality degradation due to consecutive losses. This figure for audio is 3 frames. We also observed that humans are not very sensitive to video rate variations, in contrast to the high degree of sensitivity to audio. Our results indicate that even a 20% rate variation in a newscast type video does not result is significant user dissatisfaction. The situation of audio rate variations are much more different. Even about 5% rate variation in audio is noticed by most observers. We also noticed that a momentary rate variation in the audio stream seemed amusing for a short time, but it soon resulted in being considered an annoyance, and participants concentrated more on the defect than its contents. Our results also indicate that at aggregate audio-video synchronization loss of about 20/100n human tolerance plateaus out. This figure is about 3 frames for consecutive audio-video synchronization loss. These results are consistent with the findings of [Ste96], where a constant mis-synchronization of about 120 ms. is noticed but accepted by most participants, but about 200 ms. constant mis-synchronization is considered an annoyance. Our results can be used by algorithm designers in two ways: Firstly, given a level of consumer satisfaction, they can be used to compute the maximum permissible defect of each type. Secondly, in a situation where avoidance of all types of defects is not possible, the tabulated results can be used to choose to sustain one kind of defect over any other, that results in minimal user discontent.

The rest of the paper is organized as follows. Section 2 describes our metrics for continuity and synchronization. Section 3 describes the experimental set up and methodology. Sections 4 through 7 analyses experimental results. Finally, Sec. 8 describes incidental consequences that can be drawn from our results and potential use of them, along with our ongoing work in this area. Section 9 contains a concluding summary.

2 Metrics for Continuous Media

This section reviews continuity and synchronization metrics, of which a detailed description were given in [WS96], referred therein as quality of service (QoS) metrics for continuous media.

2.1 Metrics for Continuity

Continuity of a CM stream is metrized by three components; namely *rate, drift* and *content.* For the purposes of describing these metrics, we envision a CM stream as a flow of data units (referred to as logical data units - LDU's in the uniform framework of [SB96]). The ideal rate of flow and the maximum permissible deviation from it constitute our *rate* parameters. Given the ideal rate and the beginning time of a CM stream, there is an ideal time for a given LDU to arrive/ be displayed. Given the envisioned fluid-like nature of CM streams, the appearance time of a given LDU may deviate from this ideal. Our *drift* parameters specify aggregate and consecutive non-zero drifts from these ideals, over a given number of consecutive LDU's in a stream. For eg., first four LDU's of two example streams with their expected and actual times of appearance, are shown in Fig. 1. In the first example stream, the drifts are respectively 0.0, 0.8, 0.2 and 0.2 seconds; and accordingly it has an aggregate drift of 1.2 seconds per 4 time slots, and a non-zero consecutive drift of 1.2 seconds. In the second example stream the largest consecutive non-zero drift is 0.2 seconds and the aggregate drift is 0.3 seconds per 4 time slots. The reason for a lower consecutive drift in stream 2 is that the unit drifts in it are more spread out than those in stream 1.

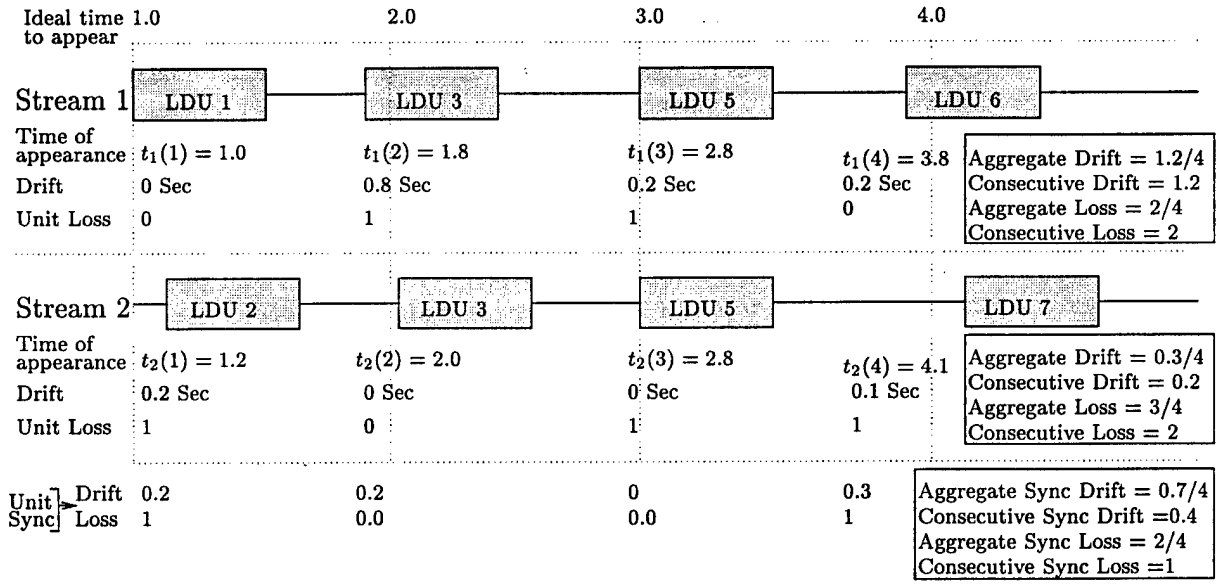


Figure 1: Two Example Streams used to Explain Metrics

In addition to timing and rate, ideal contents of a CM stream are specified by the ideal contents of each LDU. Due to loss, delivery or resource over-load problems, appearance of LDU's may deviate from this ideal, and consequently lead to discontinuity. Our metrics of continuity are designed to measure the average and bursty deviation from the ideal specification. A loss or repetition of a LDU is considered a unit loss in a CM stream. (A more precise definition is given in [WS96].) The aggregate number of such unit losses is the *aggregate loss* of a CM stream, while the largest consecutive non-zero loss is its *consecutive loss.* In the example streams of Fig. 1, stream 1 has an aggregate loss of 2/4 and a consecutive loss of 2, while stream 2 has an aggregate loss of 2/4 and

a consecutive loss of 1. The reason for the lower consecutive loss in stream 2 is that its losses are more spread-out than those of stream 1.

2.2 Metrics for Synchronization

As in the case of continuity metrics, synchronization metrics are also categorized into content, rate and drift. Rate of rendition of a collection of synchronized streams is determined by the rates (must be equal to be synchronized) of component streams. The rate variation of a collection of synchronized streams is taken to be the maximum of their component streams.

In a perfectly synchronized collection of streams, the i^{th} LDU of each stream should start playing out at the same instant of time. Failure to accomplish this ideal is measured by the maximum difference between the display start time of the LDU's in the group, and is referred to as the unit synchronization drift. The aggregate of such unit synchronization drifts over a given number of LDU slots is the aggregate synchronization drift, and the maximum of such non-zero consecutive synchronization drifts is the consecutive synchronization drift. They measure the average and bursty time drifts in synchronization. In Fig. 1, the two streams have unit synchronization drifts of 0.2, 0.2, 0.0, and 0.3 seconds respectively, resulting in an aggregate synchronization drift of 0.7/4, and a consecutive synchronization drift of 0.4 seconds.

For the content component, with streams consisting of LDU's with equal play-out times, there is a natural collection of LDU's that are to be played out simultaneously. The largest discrepancy in the LDU numbers between any two pairs in such a collection is referred to as the unit synchronization loss. The aggregate and largest non-zero consecutive unit synchronization loss is referred to as *aggregate synchronization content loss* and *consecutive synchronization content loss*, respectively. In the example of Fig. 1, due to losses of LDU's there are unit synchronization content losses at the first and the last pairs of LDU's, resulting in an aggregate synchronization content loss of 2/4 and a consecutive synchronization loss of 1.

2.3 Relationship Between Metrics

Because our continuity metrics specify permissible intra-stream deviations and losses, and synchronization metrics specify permissible inter-stream deviations and losses, there are cross effects on each other. We have been able to show that content specifications given in the form of our metrics of each individual stream are not sufficient to specify the content of synchronized streams in the same form. We have also shown that by controlling timing drifts of individual frames, it is possible to control average inter-stream timing drifts, although it is not possible to control consecutive drifts [WS96].

3 Experimental Design

Originally we planned to meaure and validate the tolerable ranges of all our metrics. Due to the inability to control timing on the computers precisely, we decided to use professionally edited pre recorded segments of audio and video. Even in the area of professional editing, our equipment was unable to control the appearance of video and corresponding audio to millisecond time granularity; the result being that we focused only on testing for content and rate parameters.

Our experiments consisted of eight groups; aggregate and consecutive content losses of audio, video and synchronization consisted six of them, and rate variations in the audio and video streams were the remaining two. Out of the eight experiments, three consisted of audio only segments,

Figure 2: Shots of Audio-Visual Clips Used in the Experiment

another three consisted of video only segments, and the remaining two consisted of audio-video segments.

3.1 Design Concerns and the Pilot Study

Several issues in survey design and psychological evaluations arise in carrying out user evaluations of human perceptions. Some of them include proper design, so that the end data can be used to test the hypothesis, and minimize the affect of extraneous variable, minimize participant biases and avoid conveying designer biases to the participants. In our design we have strived to achieve these goals. In designing our experiment, the experimental methodology and techniques of analysis used in [Ste96, AFKN94] have been useful to us.

In order to evaluate potential suitability of our experimental methodology and design to the intended task, we conducted a pilot study with about 10 participants. The results of this study and professional help [Fie96] made us change many things in the questionnaire, tapes and the environment in which the experiment was carried out. In the tape, we decided to show the two clips used in their perfect form, so that participants can judge for themselves if there were defects and if so what they were. This was due to the fact that TV and broadcast media that our participants are most familiar with do not usually have the kind of defects that we may want observed, and also to provide a baseline for comparisons. We provided a potential list of defects, some of which were not in our clips. This was due a suggestion that many participants may not find common words to describe a defect, and at the end categorizing defects would lead to too many categories. We decided to categorize defects of the same type with an error free clip included, but unindentifed in to a randomly arranged succession of clips. Each section containing audio, video or both were identified as such, so to avoid the absence of either media type to be considered a defect.

In the design of the survey, we had to make substantial changes after the pilot study. It was determined that the survey will be in the tabulated format, as opposed to having a page per clip, as the sheer size of the survey form seem to discourage some potential participants. The order and wording of questions had to be changed to suit an average American college going audience. We also decided not to allow individuals to take the survey on their own, so that the environment of the presentation, and answers to participant doubts and questions during the experimental runs remain constant. The Likert scale was changed from [1, 8] to [1, 10], where 1 way poor and 10 was excellent. We also asked the participants to categorize each clip as *Do not mind the defect if there*

Experiment	Media	Defect in Test Clips					
Aggregate Loss	Video	6/100	21/100	12/100	3/100	0/100	
Consecutive Loss	Video	0	1	5	4	3	2
Rate Variation	Video	10%	0%	02%	20%	15%	6%
Aggregate Loss	Audio	6/100	21/100	12/100	3/100	0/100	
Consecutive Loss	Audio	0	1	5	4	3	2
Rate Variation	Audio	10%	0%	02%	20%	15%	6%
Aggregate Synchronization Loss	A/V	40/100	4/100	16/100	24/100	0/100	
Consecutive Synchronization Loss	A/V	15	3	10	0	5	20

Table 1: Order of Defects in Test Clips

is one, *I dislike it and its annoying*, and *I am not sure* similar to the survey in [Ste96].

3.2 Design Decisions

Thirty second audio-video segments were taken from a bust view of two articulate speakers (Fig 2), with no particular accents, describing neutral subjects. The chosen speakers were unknown to participants in the study. This was done to avoid any biases that may carry over about the speakers into our study. Neutral accents were chosen to avoid any mis-interpretation of words in the face of introduced defects, and also to give our participants the benefit of listening to a voice that comes with the most familiar pronunciation. The contents of the two speakers were about what care they would take in organizing their lectures, and about concentration spans of junior high school students. None of our participants were teachers, nor junior high students. The length of test segments were chosen to be 20 to 30 seconds because according to [Ste96] about 20 seconds suffices for participants in a MM user study to form their opinions about a clip. Although the head view results in the most number of defects being perceived [Ste96], we chose the bust view because it represents the news media type of a situation better than a talking head occupying an entire screen.

3.3 Parameters Used in Experiments

The tapes were made with the following characteristics. In the aggregate media loss experiments, the consecutive losses were kept to a constant of 3 video frames, under the normal speed of 30 frames per second. The media losses were created by introducing *jump cuts* in the NTSC time code. For the rate variation experiment, a nominal rate of 30 frames per second rate was maintained, but a square sinusoidal wave with each quarter wave lasting 5-6 seconds was produced. For the aggregate synchronization loss experiment the consecutive synchronization loss was kept to 4 video frames at 30 frames/second speed. For the consecutive synchronization loss experiment the aggregate synchronization losses were kept to 40/100. The master tape consisted of an introductory part lasting about 3 minutes, after which the two perfect clips were shown, followed by three groups of experiments: video, audio and synchronization. Within each group, sub-group order was aggregate loss, consecutive loss and rate variation experiments. Within each experiment, defective clips were arranged in the random order given in Table 1. For each experiment there were about 5 to 6 clips, with varying degrees of controlled defects, that were shown in random order.

Experiments with Video Only Clips

These experiments have NO SOUND. Please watch the silent video and fill out the following tables.

Clip Number	Grade the quality of the clip	Did you notice a defect ? If so, please describe it	If your TV programs had this error how would you categorize it?		
	1 (poor) to 10 (excellent)	i.e. skip, stutter breaks, mis-synchronization, gaps distortions etc.	I don't mind the defect	I dislike it. its annoying	I am not sure It depends
Group 1 Clip 1	1 2 3 4 5 6 7 8 9 10				
Clip 2	1 2 3 4 5 6 7 8 9 10				
Clip 3	1 2 3 4 5 6 7 8 9 10				
Clip 4	1 2 3 4 5 6 7 8 9 10				
Clip 5	1 2 3 4 5 6 7 8 9 10				

Figure 3: A Sample Table from a Blank Survey Form

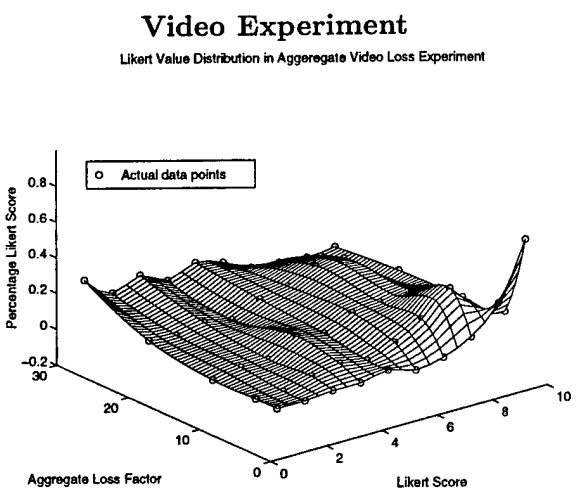
3.4 Administering the Experiment

Experiments were conducted in small groups of 3 to 6 participants chosen mostly from students at the University of Minnesota, who participated in our study voluntarily. In order to draw participant attention to potential defects, the background noise was kept to a minimum and the contents of clips were deliberately made to be boring. We also told the participants that the objective of our study was to look for defects, and provided a sample list of them. At the beginning of the survey we showed the two clips in their perfect form. As expected, most participants found the endeavor boring and very repetitive, although a fair number found some clips to be rather amusing. All eight groups were shown in one sitting that lasted about 45 minutes. After each clip was shown, the participants were asked to fill out the corresponding row of scores in a survey form. The sample survey used for the first clip is given in Fig. 3. The survey consists of an introductory description, six tables (one per each experiment) and a questionnaire about participants experience with TV production. As seen from the sample table given in Fig. 3, each participant had to grade each clip on a Likert scale [Opp83] from 1 to 10, give possible defects detected, and state if the defect was annoying, not so, or could not decide, which we call the *acceptability score.*

3.5 Processing the Surveys

The results of the surveys were entered into a database, and visualized using Matlab [PESMI96]. As expected, increase in defects resulted in a decrease of user satisfaction, except for the experiment on aggregate losses of audio. The data as, taken from the surveys, and average and standard deviations of Likert values, and the ratio of participants who considered the clip to be perfect, acceptable and unacceptable were graphed per each experiment. These garphs were smoothed by using a cubic spline interporlation provided by Matlab. The analysis of the data and conclusions drawn from them follow in Sections 4 through 7.

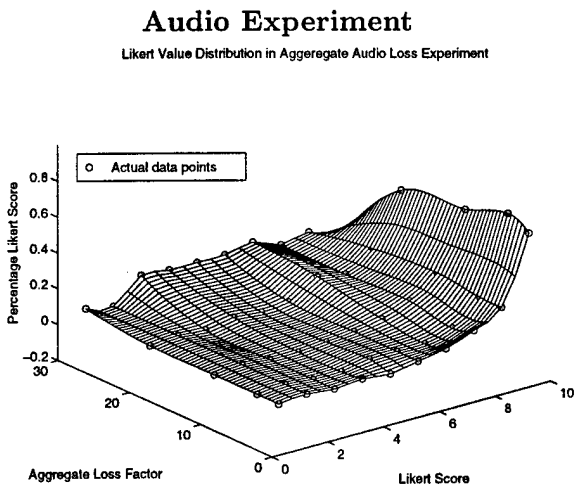
Two remarkable trends emerge from our results: One is that there are defects for which there is a gradual increase in user discontent with increasing defects. Aggregate video loss is a clear example of this kind. The other is that there is a sharp increase of user discontent that plateaus out after



A

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
3/100	0	0	1	1	3	0	2	7	17	39
6/100	0	0	3	5	3	7	13	22	10	7
12/100	0	2	2	8	6	11	12	14	10	5
21/100	4	3	5	3	8	10	15	15	3	4
30/100	16	8	12	7	11	8	3	2	1	2

B



C

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
3	0	1	0	1	0	2	4	8	14	40
6	0	1	0	0	2	2	2	5	14	44
12	0	1	0	0	1	2	4	8	16	38
21	0	1	0	2	1	2	9	7	14	34
30	3	1	10	9	9	9	11	7	9	2

D

Figure 4: Data from the Aggregate Loss Factor Experiment

a specific value. Synchronization and consecutive losses are clear examples of this kind. Rate fluctuations are some where inbetween, and for certainity, humans seemed to be far less tolerent to audio rate fluctuatins than to that of video.

4 Aggregate Loss Experiment for Media Streams

As stated, there were five clips with aggregate media losses ranging from 3/100 to 21/100, with a consecutive loss factor of 3 LDU's. The order of these clips were arranged as given in Table 1. For the aggregate loss experiment of video streams, as evident from data tabulated in Fig 4 (B)and visualized in Fig 4 (A), as the aggregate media loss increases, the distribution of Likert values shift from the higher end towards the lower end of the spectrum. This trend indicates that increased aggregate video loss leads to increased viewer discontent.

We were expecting the same trend in the corresponding experiment on audio, but as observed from data tabulated in Fig.4 (D) and visualized in Fig.4 (C), our expectations were not fulfilled to the same extent as for video. A closer examination of our tapes reveal that most eliminated LDU's in the audio stream correspond to silence. Consequently, although it requires further experiments to justify our speculation about aggregate audio drops, current results indicate that aggregate silence elimination in the audio stream does not result in considerable user discontent in the range from 0/100 to 21/100. We speculate that further silence elimination would reach the point of considerable listener discontent. Notice that the higher end Likert scales of Fig. 5 (D) provide evidence in support of this trend. Our ongoing work includes further experimentation to settle this speculation.

To further our understanding of the pattern of user discontent, we tabulated and visualized the average and standard deviations of Likert values against the losses for video and audio, given in Fig. 5 (A)and (C) respectively, which clearly brings out the trend. The lower standard deviation at

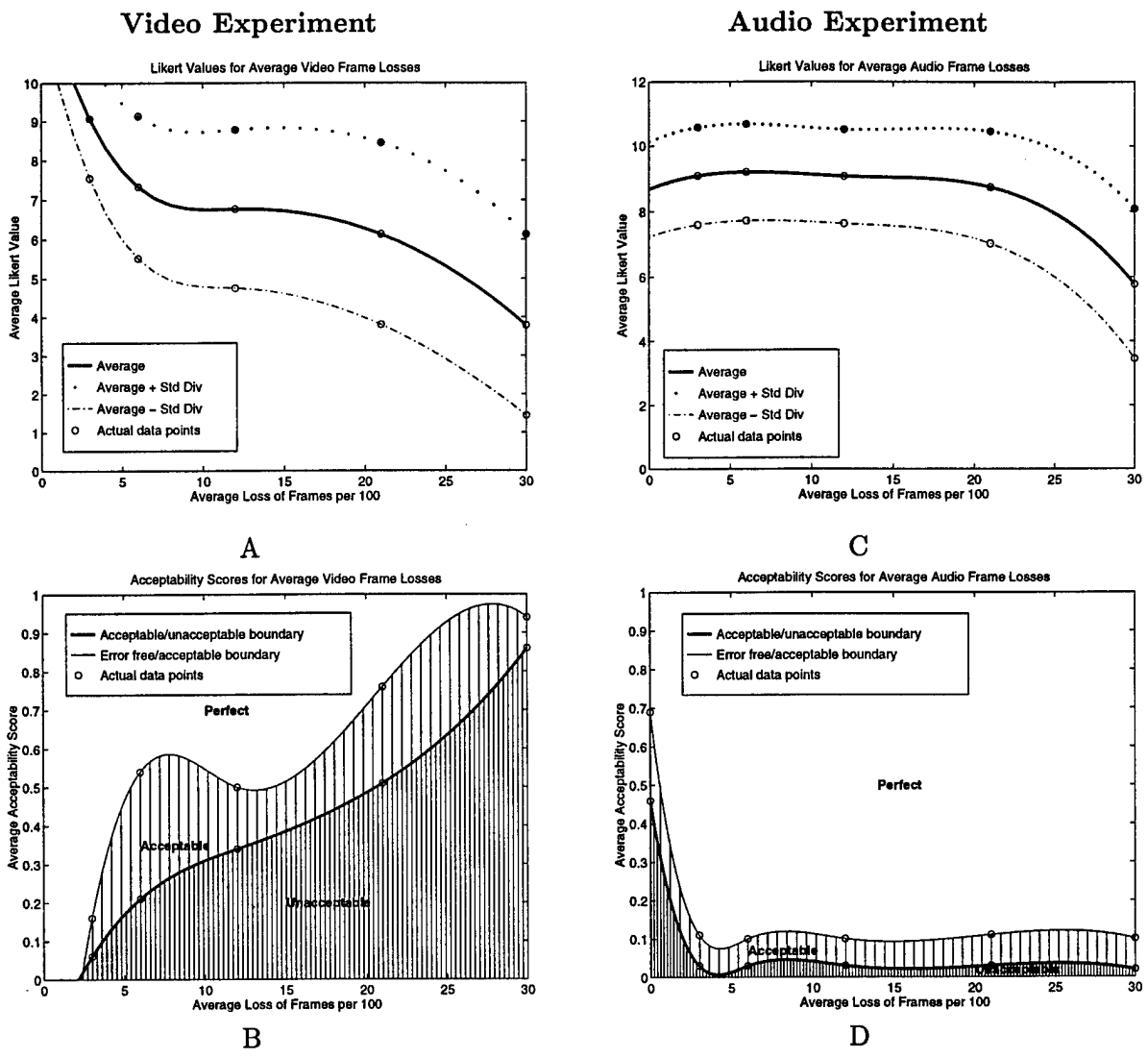


Figure 5: Summarized Results of the Aggregate Loss Factor Experiment

the higher values of the average Likert scale indicates that there is higher consensus in the judgment expressed by its mean. Also notice that the maximum standard deviation is about 2, a reasonable 1/5 of the total score.

The acceptability scale, tabulated and graphed in Fig. 5 (A) and (C) respectively, yields the regions in which the users expressed clear intolerability, the willingness to tolerate and the perfect acceptance. In all the graphs we have, we notice a correlation between the average Likert value in the Likert scale and the curve that separates the *unacceptable* region from the rest. This seems to indicate that the two metrics that were used in two other reported user studies in multimedia [Ste96, AFKN94], namely the Likert and the acceptability scales have a strong relationship to each other, and consequently can be used in our type of study interchangeably.

If the Likert and acceptability scores are graphed together, the former intersects the acceptable curve in the latter at about 17/100 aggregate media loss and the unacceptable curve at about 23/100 media losses. Modulo our experimental results, these observations imply that 17/100 to 23/100 is the noticeable but tolerable region for aggregate video losses. Similar analysis applied to the results of the audio experiment yields that within our operational range (i.e 0/100 to 21/100) aggregate audio losses were unnoticed.

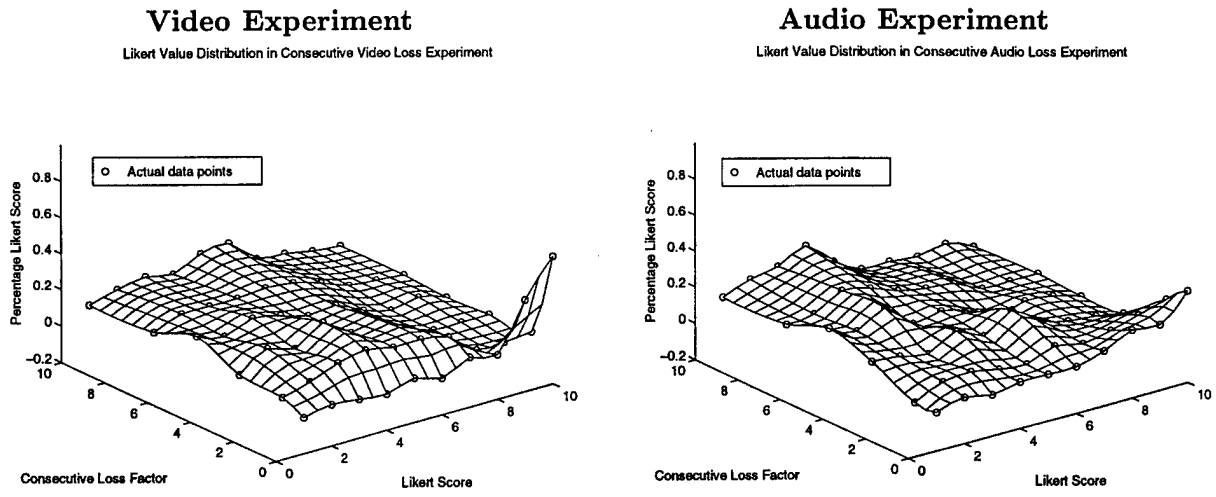
5 Consecutive Loss Experiment for Media Streams

There are six clips with aggregate media losses ranging from 0 to 10 consecutive LDU's, with the order of arrangements within the experiment as given in Table 1. As seen from results tabulated in Fig. 6 (B) & (D) and visualized in Fig 6 (A) and (C), increasing consecutive losses result in a sharp rise in viewer discontent. This is evidenced by the concentration of lower Likert values around 3 to 5 consecutive media losses in data from both video and audio streams, as given in Fig. 6 (B) and (D), respectively.

This trend is further clarified by the average Likert and acceptability graphs given in Fig. 7 (A),(C) and Fig. 7 (B),(D) respectively. As seen in Fig. 7 (D), for audio streams 3 to 4 consecutive frame losses receive an Likert score of 9. For video, as seen from Fig. 7 (B) this limit is 2. Compared with video aggregate loss experiments visualized in Fig. 5, acceptability scores have a thin margin for noticeable but tolerable consecutive losses, although the margin for video losses is slightly higher than those for audio. In contrast to average video losses, graphed in Fig. 5 (B), user discontent with consecutive losses sharply rises and then plateaus out at 2 and 3 frames for video and audio respectively. Standard deviation for acceptability values for both media, as visualized in Fig 7 (A) and (C) is approximately 2 units. At the high end of the scale the standard deviation for the video stream is lower, indicating more consensus in the rating. Because of the thin margin for the acceptable region, the intersection of Likert graphs and acceptability graphs remain single values of 1 for video and 2 for audio.

6 Rate Variation Experiment

As stated, there are six clips with 0% to 20 % rate variation from an average of 30 frames/seconds with a pattern of a square sinosidel curve of five frame quarter length. The order of these clips were arranged as given in Table 1. As evident from data tabulated in Figs. 8 (B), (D) and visualized in Figs. 8 (A), (C), user discontent shifts from the higher end towards the lower end of the spectrum with the increase in the amplitude of the sinosidel rate wave, indicating that increasing rate fluctuations lead to increasing viewer discontent.



A

Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	2	1	0	3	0	5	3	21	35
1	4	7	11	13	11	11	9	1	1	2
3	5	12	9	16	9	11	5	1	1	1
5	12	9	13	11	8	9	4	1	3	0
7	6	10	10	10	11	10	7	3	2	1
10	5	8	10	8	13	14	5	4	2	1

B

C

Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	1	4	2	4	4	4	7	12	11	21
1	1	6	7	5	4	8	6	8	11	14
3	9	8	6	12	7	14	7	4	3	1
5	14	10	16	7	9	6	1	4	3	0
7	8	9	15	13	9	7	6	1	1	1
10	7	11	13	18	9	3	3	2	4	0

D

Figure 6: Data from the Consecutive Loss Factor Experiment

To further our understanding of the pattern of user discontent, we tabulated the average and standard deviations of Likert values against the losses, as given in Figs. 9 (A), (C), which clearly brings out the trend. The lower standard deviation at the higher values of the average Likert scale indicates that there is higher consensus in the judgment expressed by its mean. Also, the maximum standard deviation in Figs. 9 (A), (C) is about 2. Notice that the average Likert value in the audio decreases more uniformly, compared to video. This trend implies that we are not very sensitive to the rate fluctuations in video, as compared to those of audio. Also, audio has a uniformly higher Likert value than video, further substantiating this stance. Data on acceptability scores have been visualized in Figs. 9 (B), (D), and shows the corresponding plateaus and trends as those in average Likert scales.

If the Likert and acceptability scores are graphed together, the former intersects the latter for video at about 7% and 8%. These results imply that upto about 20% of video and 7% of audio rate variations are tolerated, and after about 8% audio rate variations become intolerable. In this experiment two metrics, namely average Likert and average acceptability scales show a strong positive co-relation.

7 Transient Synchronization Loss Experiments

As stated, there are six clips each for aggregate and synchronization loss experiments. In the aggregate loss experiment they range from 0/100 to 40/100 with a constant consecutive loss of 4, and in the consecutive losses experiment they range from 0 to 20 with a aggregate synchronization loss of 40/100. The order of these clips were arranged as given in Table 1. For synchronization loss experiments, as evident from tabulated data in Figs 10 (B), (D) visualized in Figs 10 (A), (C), as the losses increase, the distribution of Likert values shifts from the higher end to the lower end of

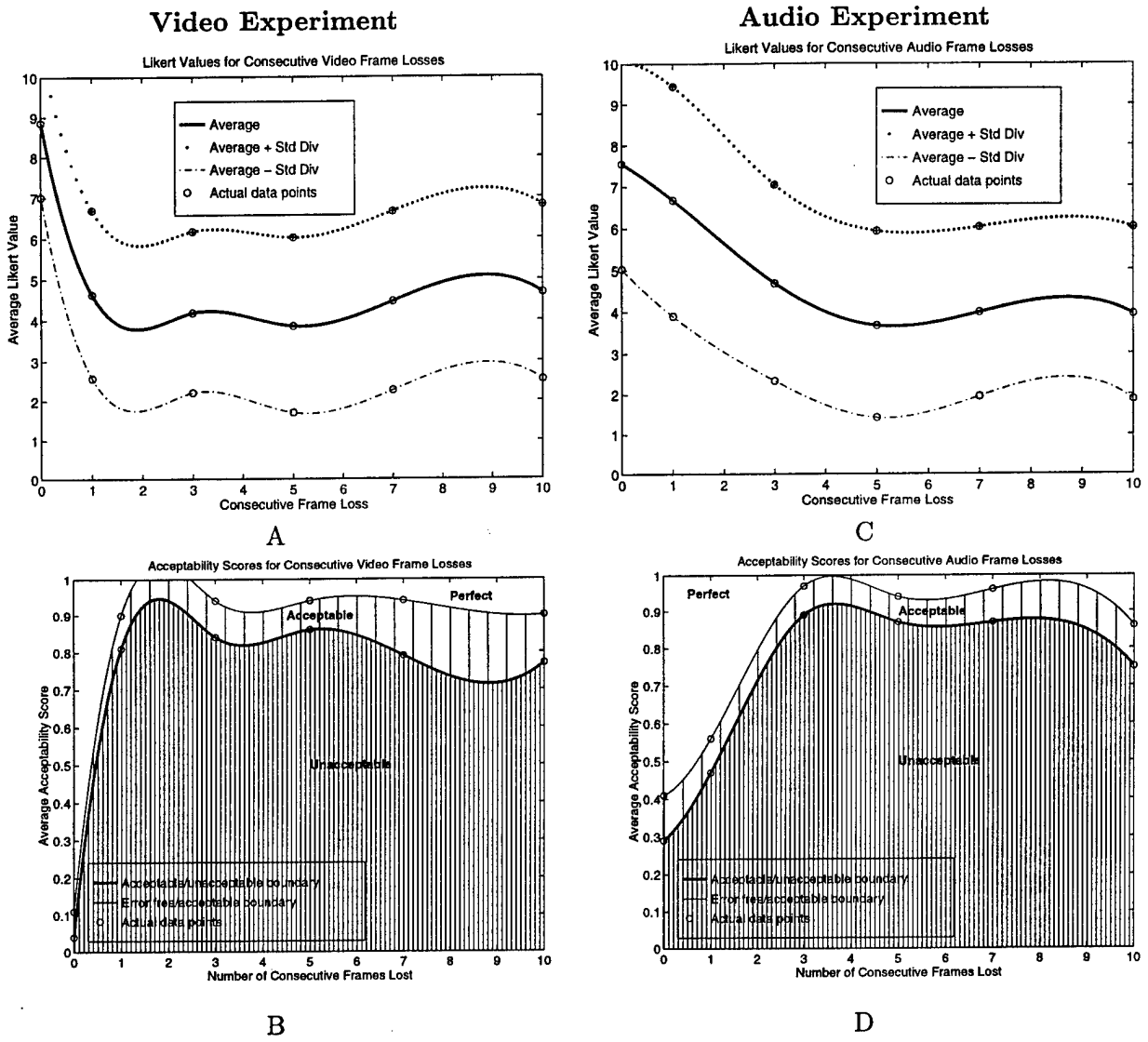
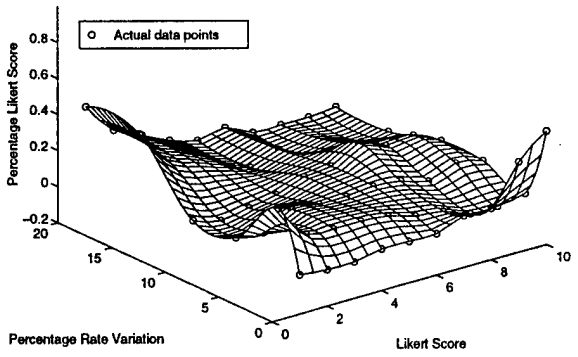


Figure 7: Summarized Results of the Consecutive Loss Factor Experiment

Video Experiment

Likert Value Distribution in Audio Rate Variation Experiment



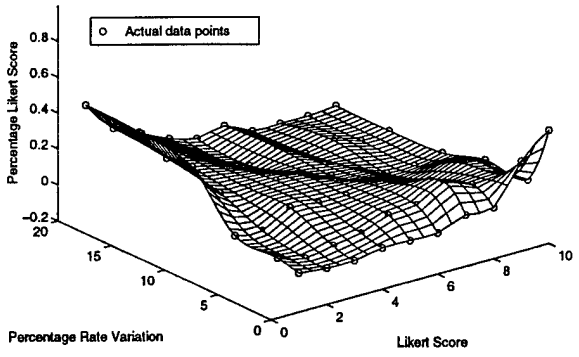
A

% Rate Variation	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	2	0	1	1	2	1	5	10	18	30
2	0	1	4	4	7	11	11	10	10	12
6	0	1	1	1	8	6	6	13	20	14
10	0	1	2	2	6	5	10	11	17	16
15	1	1	4	2	7	8	8	12	13	14
20	0	2	5	5	6	12	9	13	9	9

B

Audio Experiment

Likert Value Distribution in Audio Rate Variation Experiment



C

% Rate Variation	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	1	0	0	2	2	2	6	6	21	30
2	3	0	2	4	8	9	15	11	12	7
6	4	3	9	8	8	9	8	5	9	7
10	23	10	9	9	8	5	4	0	0	2
15	27	14	10	6	2	6	1	3	0	1
20	28	16	11	6	3	5	0	0	0	1

D

Figure 8: Data from Rate Change Experiment

the spectrum, indicating that increased transient synchronization losses lead to increased viewer discontent.

To further our understanding of the pattern of user discontent, we tabulated the average and standard deviations of Likert values against the losses given in Fig. 11 (A), (C), which clearly brings out the trend of average Likert score decreasing with increasing synchronization losses. As in the case of consecutive media loss experiments, these have sharp increases in their acceptability scale and plateaus out around 12/100 and 3 for average and consecutive losses respectively.

The acceptability scale, visualized in Figs. 11, (B), (D), give the regions in which the users expressed clear intolerability, the willingness to tolerate, and perfect acceptance. This scale also sharply decreases and plateaus out at 12/100 and 3 for average and consecutive losses.

The intersections for average Likert and acceptability curves indicate that 6/100 to 7/100 is the range for tolerable average synchronization losses, one frame is the tolerability limit for consecutive synchronization losses.

As like in all other graphs, we notice a clear correlation between the average Likert value and the curve that separates the *unacceptable* region from the rest in the acceptability scale, indicating a strong relationship between them in synchronization experiments.

8 Further Inferences and Usage of Experimental Results

This section provides some further inferences from our experimental data, their projected usefulness and our ongoing work in this area.

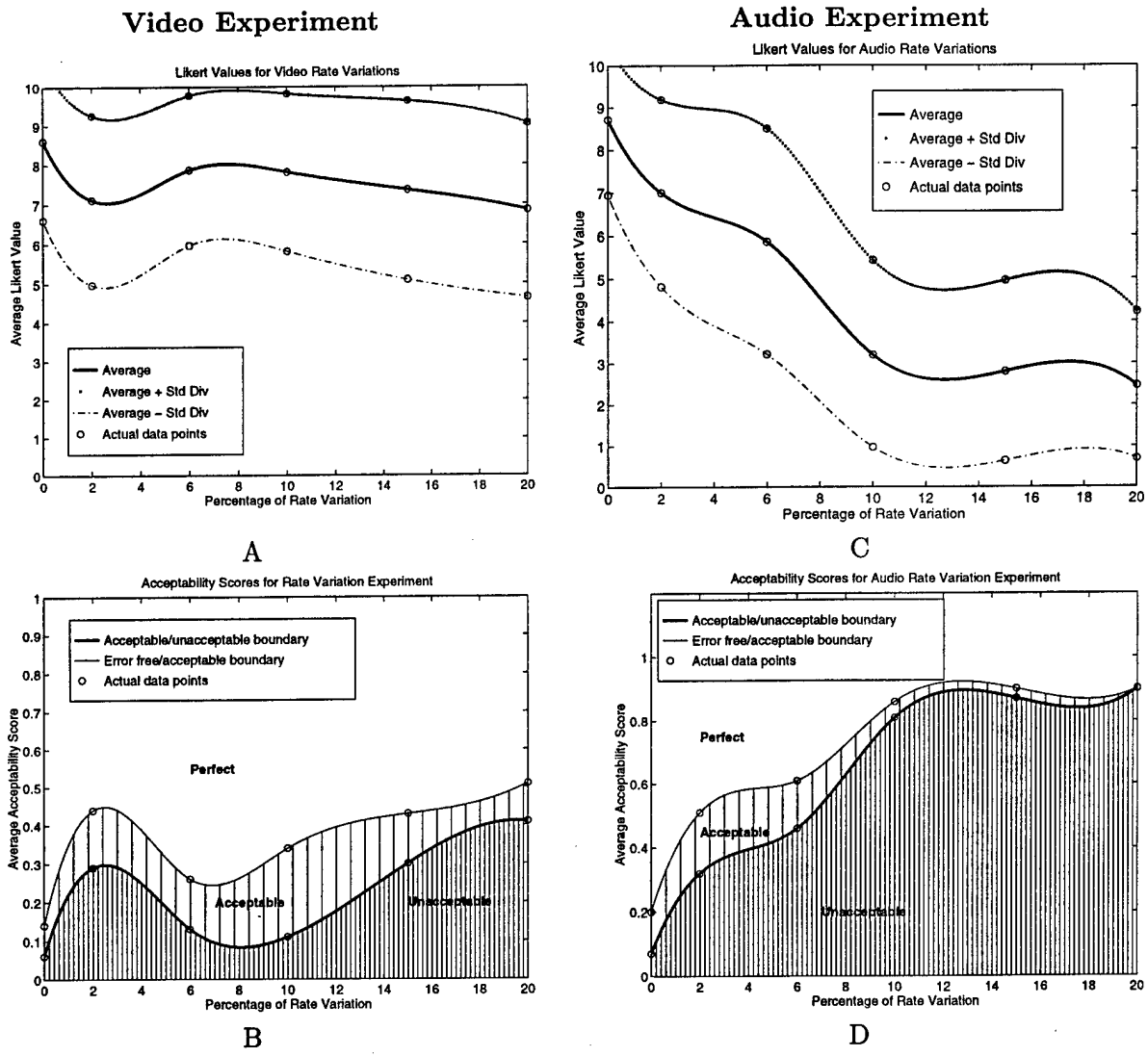
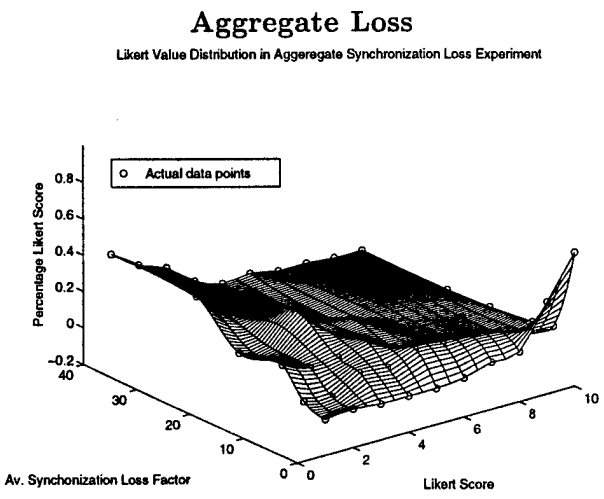


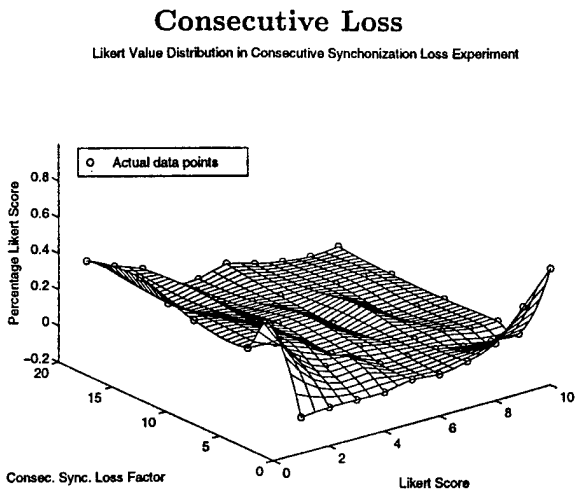
Figure 9: Summarized Results of the Fluctuating Rates Experiment



A

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	1	0	0	0	1	4	5	21	37
4	3	2	8	7	11	10	10	9	6	5
8	13	9	10	13	8	8	3	2	1	3
16	10	10	20	9	11	6	2	1	0	1
24	24	15	12	10	4	0	2	2	1	0
40	25	18	14	6	2	3	1	1	0	0

B



C

Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	1	1	1	3	2	4	8	19	31
3	30	14	8	8	3	3	3	1	0	0
5	17	16	13	6	6	4	3	3	1	1
10	18	16	12	8	8	2	6	0	0	0
15	25	12	14	5	4	3	3	2	2	0
20	22	17	13	4	3	6	3	1	0	1

D

Figure 10: Data from Synchronization Loss Experiments

8.1 Further Inferences from Experimental Results

As stated, two remarkable trends emerge from our results: One is that there are defects for which there is a gradual increase in user discontent with increasing defects. Aggregate video loss is a clear example of this kind. The other is that there is a sharp increase of user discontent that plateaus out after a specific value. Synchronization and consecutive losses are clear examples of this kind. Rate fluctuations are some where in-between, and for certainty, humans seemed to be far less tolerant to audio rate fluctuations than to that of video. Based on our observations, although we concur with synchronization experimental results obtained in [Ste96], we speculate that not all QoS experiments are going to result in such clear cut boundaries for distinguishability, tolerance and unacceptability for QoS metrics, but they gradually decrease throughout a continuous spectra of values. This trend is clearly evidenced in our aggregate loss experiment for video, and also in the rate experiments of [AFKN94].

In addition to determining the acceptable ranges for some of our QoS parameters, we can also determine their relative importance. For example, we can directly compare the Likert values of aggregate video losses and aggregate synchronization losses to determine the loss ranges where one of them is more crucial than the other. Some of the potential benefits of these are discussed in Sec. 8.2.

8.2 Use of Experimental Results

Our findings can be used in multimedia testbed designs in two different ways. Firstly, given a Likert value, or an acceptability score that would characterize the required *degree of user satisfaction*, it is possible to determine the tolerance to a given defect. For example, with a 0.8 Likert value a video stream can sustain 20/100 average loss, 1 consecutive loss, and upto 20% rate variation. For

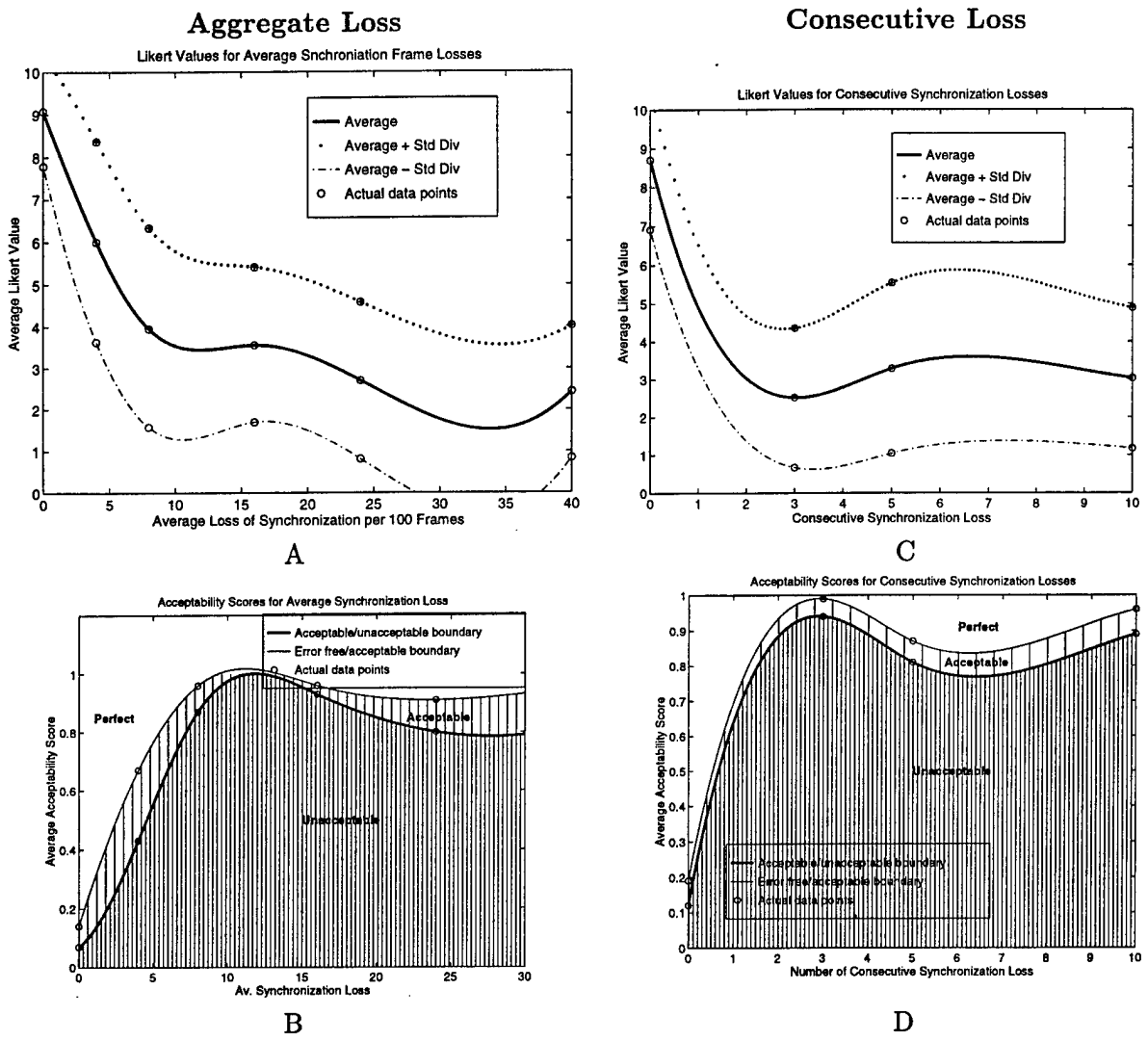


Figure 11: Summarized Results of Synchronization Loss Experiments

the audio stream these parameters are 30/100 aggregate silence elimination, 0.7 seconds worth of consecutive sample losses, and about 10% rate variation. For audio-video synchronization, they are about 7/100 aggregate losses and one consecutive loss. For a given level of user satisfaction, the tolerances of a set of defects, such as the media and synchronization losses investigated in the present paper, can be used directly as limiting values for the corresponding defects. For example, for 80% user satisfaction, we may have 20/100 as the maximum permissible aggregate video loss.

Secondly, in designing algorithms, we can assign relative weights to these losses. For example, comparing the average Likert values of video loss with consecutive synchronization losses, it is clear that the unacceptability region for the former is below that of the latter, and therefore, dropping video frames on the average is preferred to loosing synchronization consecutively. To compute relative weights for different parameters, we may assign them weights proportional to the average of some user preference parameter such as the average of all Likert value assigned for that parameter, that can be achieved for the given testbed. For example, if a designed testbed can only deliver with a aggregate video loss of 10/100, and a consecutive synchronization loss of 5, compute the average of the Likert values over [0, 10/100] for the aggregate video loss and [0, 6] for the consecutive synchronization loss. Suppose that the former is 7 and the latter is 5.5, then assign these as weights of importance, during dynamic play-out management. A potential usage of such weights is that the parameter that carries the smallest weight in the range of operation can be sacrificed in order to avoiding defaulting on ones with higher weights.

8.3 Comparison with Existing Work

Parameters of human tolerance to audio-video and audio-pointer synchronization were obtained in [Ste96]. They were categorized as undetected, detected but tolerable, and intolerable errors. These parameters are for loss-less streams. In a CM client-server paradigm, streams may be delivered through a network. At the lower levels of the protocol stack, the network can drop packets, and in order to recover the losses, some kind of retransmission is necessary. This may induces intolerable delays and jitters in the CM stream. Suppose instead, the application itself allows for a lossy media stream, through some QoS based loss characteristics of CM streams, then the retransmission may be unnecessary, and consequently, the delay and jitter at the application level and bandwidth at the network level can be saved. Our parameters can be used to compute such QoS based LDU drops at the application level.

Another observation we had was that in our testbed, audio and video drifted in and out of synchronization, as opposed to being static values. Granted that if maximum drifts were within limits reported in [Ste96], then the static limits stated therein would apply; but we were speculating that for transient mis-synchronizations, the participants would be more forgiving. As the reported data indicates, this is not the case.

[AFKN94], categorizes audio-visual clips as *high* and *low* in audio, video and temporal dimensions, referred to theirin as *video classification schemas (VCS)*. They measure the perceptual importance of each dimension in conveying the total message contained in clips across to the intended viewer. For example, sports footage and talk shows are considered high and low in the temporal dimension, respectively. Such a classification, while rich in semantics and its relevance to human perception, requires some extra efforts and the servers need to be enriched to understand their relevance, and extra efforts by the producers or some other intermediate personnel. In this respect our test clips should be considered low in the temporal dimension and (perhaps) video dimension, but high in audio dimension. The reported study categorizes the effect of play-out rates on audio-visual demonstrations with different VCS schema values. This study, while important, does not cover the loss parameters, transient mis synchronizations, and rate fluctuations, all of which

can happen during audio-visual display. The Likert scores of [AFKN94] is from 1 to 7, whereas our scale is from 1 to 10. In addition, we also use the scale of [Ste96]. One of the advantages of this study is the block design of the experiment in which the combined effect of multiple parameters variations on perception were determined, whereas in our experiment, we have only determined the effects of individual parameters.

8.4 Limitations of the Current Experiment and Our Ongoing Work

The aggregate loss experiment for video needs to be re-done with appropriate clips, due to the reason that we eliminated silence rather than voice. We are also in the process of comparing our results with known perceptual studies of silence elimination. Another parameter we would like to know is the perceptual difference in skipping video frames, and repeating the same frame., because our stream loss metrics treat them as equal.

Secondly, we would like to get the combined effect of our parameters on human perception. In this respect, combining our results with those of other studies to obtain a combined Likert scale as a function with multiple inputs as defects will be most beneficial. We are also planning a block-designed factorial [Edw85] experiment involving more QoS parameters. As stated, this involves having a sufficiently randomized experiment where the participants boredom does not effect their judgment. Some of our ongoing work addresses this issue in detail. The benefit of such a study are significant in the implementation of multimedia test-beds, as given below.

- It allow to prioritize user needs.
- It allows for the most beneficial dynamic QoS adjustments [AFKN94].
- It adds up to building a comprehensive user level QoS metric for multimedia [Sta96].
- It helps in resource management [Sta96].
- It helps in exception handling and fault tolerance [Nai96].
- It can be used in multimedia server design [LSSKH97].

We are also in the process of enhancing the Tcl/Tk [Wel95, Ous94] based Berkeley Continuous Media Toolkit (CMT) [SRY93] to enhance its performance by using our new-found tolerances to defects reported in this paper. In that work we see a clear need for a comprehensive QoS metric.

9 Conclusions

Based on the simple observation that (1) loss of media content, (2) rate variations and (3) the degree of transient mis-synchronizations result in user discontent in multimedia demonstrations, we designed metrics to measure these phenomenon. A user survey to substantiates our initial observations, and thereby validating the assumptions that underly our model. Use of our experimental results in multimedia algorithm design has been discussed.

References

[AFKN94] R.T. Aptekar, J.A. Fisher, V. Kisimov, and H. Nieshlos. Distributed Multimedia:User Perception and Dynamic QoS. In *SPIE* , volume 2188. SPIE, 1994.

[Edw85] Allen E. Edwards. *Experimental Design In Psychological Research.* Harper and Row, 5 edition, 1985.

[Fie96] Molly Fiedler. personal communication, December 1996.

[Geo96] Nicolas D. Georganas. Synchronization issues in multimedia presentational and conversational applications. In *Proceedings of the 1996 Pacific Workshop on Distributed Multimedia Systems (DMS'96)*, page invited talk, June 1996.

[HRKHS96] Jiandong Huang, Jim Richardson, Deepak R. Kenchamanna-Hosekote, and Jaideep Srivastava. Presto: Final technical report. Technical report, Honeywell Technology Center, august 1996.

[LSSKH97] Wonjun Lee, Difu Su, Jaideep Srivastava, and Deepak R. Kenchamanna-Hosekote. Qos driven scheduling for cm file servers. In *SPIE97, VV4*, November 1997.

[Nai96] Kshirasagar Naik. Exception handling and fault-tolerence in multimedia synchronization. *IEEE Journal on Selected Areas in Communication*, 14(1):196–211, 1996.

[Opp83] A. N. Oppenheim. *Questionnaire Design and Attitude Measurement.* Heinemann, 1983.

[Ous94] John K. Ousterhout. *Tcl and the Tk Toolkit.* Addison-Wesley, 1994.

[PESMI96] Eva Part-Enander, Andres Sjoberg, Bo Melin, and Prenilla Isaksson. *The Matlab Handbook.* Addison-Wesley, 1996.

[SB96] Ralf Steinmetz and Gerold Blakowski. A media synchronization survey: Reference model, specification and case studies. *IEEE Journal on Selected Areas in Communication*, 14(1):5–35, 1996.

[SGN96] Ralf Steinmetz, Nicolas D. Georganas, and Toru Nakagawa. Guest editorial: Synchronization issues in multimedia communications. *IEEE Journal on Selected Areas in Communication*, 14(1):1–4, January 1996.

[SRY93] Brian Smith, Larry Rowe, and S. Yen. A Tcl/Tk Continuous Media Player. In *Proceedings Tcl 1993 Workshop*, June 1993.

[Sta96] Richard Alan Staehli. *Quality of Service Specification for Resource Management in Multimedia.* PhD thesis, Oregeon Graduate Institute of Science and Technilogy, January 1996.

[Ste96] Ralf Steinmetz. Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communication*, 14(1):61–72, January 1996.

[Wel95] Brent Welch. *Practical Programming in Tcl and Tk.* Prentice Hall, 1995.

[WS96] Duminda Wijesekera and Jaideep Srivastava. Quality of Service (QoS) Metrics for Continuous Media. *Multimedia Tools and Applications*, 3(1):127–166, September 1996.